

# Detecting hitchhiking from patterns of DNA polymorphism

Justin C. Fay

Department of Genome Sciences,  
Lawrence Berkeley National Laboratory,  
Berkeley, CA, 94702

Chung-I Wu

Department of Ecology and Evolution,  
University of Chicago,  
Chicago, IL, 60637

The genetic basis of adaptive evolution has long escaped the grasp of evolutionary geneticists due to the difficulty of mapping an organism's phenotype to its genotype. However, adaptive substitutions may also be identified by their effects on linked neutral variation. This has made it possible to test whether an adaptive substitution has recently occurred in a particular gene and whether such substitutions are common within an organism's genome. Of critical importance is the power of tests that detect adaptive substitutions and our confidence in the evidence for such events.

Adaptive substitution can be detected by their effects on levels and patterns of DNA polymorphism. With few exceptions all tests compare some feature of observed polymorphism data with that expected under a Wright-Fisher neutral model. This model assumes mutations arise in a diploid population of size  $N$  with probability  $\mu$  per generation, mating is random, there is no selection, there is no population structure, population size is constant, there are non-overlapping generations, and the population is at mutation-drift equilibrium [13]. Although it is true that natural populations violate most of these assumptions, the neutral model is often sufficient to describe most features of polymorphism data obtained from natural populations. This

is in part due to the fact that slight violations of these assumptions do not cause large deviations from the neutral expectation and in part because under neutrality nearly all features of polymorphism data are expected to be quite variable.

In this chapter we describe how various aspects of polymorphism data can be used to detect the effect of positive selection on linked neutral variation, or the hitchhiking effect. We also compare these methods, with respect to their power to detect hitchhiking and their sensitivity to violations of the Wright-Fisher model.

## Reduction in levels of variation

The primary effect of positive selection on linked neutral variation is a reduction in heterozygosity (Figure 1). In the absence of recombination, variation is steadily removed by hitchhiking or the spread of an advantageous allele through a population. Subsequent to hitchhiking variation is slowly regained by the drift of new mutations to detectable frequencies. When selection is strong the advantageous allele is fixed in approximately  $\ln(2N)(2/s)$  generations, compared to a neutral allele which is expected to take  $4N$  generations, where  $N$  is the effective population size and  $1/2N$  is the initial frequency of the advantageous mutation [32]. Subsequent to a hitchhiking event most variation is regained within  $4N$  generations [55] [43].

In the presence of recombination, the reduction in heterozygosity is a function of the ratio of the rate of recombination to the selection coefficient,  $c/s$ , and the initial frequency of the advantageous mutation, assuming the spread of the advantageous mutation is deterministic [36]. This assumption is justified when the frequency of an advantageous mutation is greater than  $\varepsilon$  but less than  $1 - \varepsilon$ , where  $\varepsilon$  is the frequency at which the probability the advantageous mutation is lost is nearly zero, i.e.  $(1 - 2s)^{2N\varepsilon} \approx e^{-4Ns\varepsilon} \approx 0$ , where 1,  $1 + s$  and  $1 + 2s$  are the fitnesses of genotypes aa, Aa and AA, respectively [29]. Various approximations have been made to account for the hitchhiking dynamics below  $\varepsilon$  and above  $1 - \varepsilon$  [29] [5] [45] [6], but if selection is strong, the stochastic phase of the hitchhiking event does not much influence the time to fixation [5]. However, it should be noted that recombination events that occur when the advantageous mutation is rare can have a large effect on the reduction in heterozygosity at a nearby locus. Thus, even a slight change in the time spent between  $1/2N$  and  $\varepsilon$  is expected to magnify

or reduce the effects of recombination on hitchhiking [5].

A reduction in heterozygosity can be used as evidence for hitchhiking. The HKA test [26] detects a reduction in heterozygosity at one locus compared to a reference locus and has been applied to many genes in *Drosophila melanogaster* [37]. Although the test accounts for different mutation rates at different loci within the genome, the test can be difficult to interpret since the significance of the test varies depending on which "neutral" locus is used as a reference. The HKA test is also sensitive to population subdivision, which increases the variance in heterozygosity across the genome [51], and purifying selection which is expected to reduce levels of variation as a function of the rate recombination and the rate of deleterious mutations [10]. More compelling arguments for hitchhiking can be made by showing a local reduction in variation along a chromosome (as shown in Figure 1). This has been done for the *Acp26Aa* [1] [15], *Sod* [25] [27] and *Sdic* genes [38] in *D. melanogaster*. However, even under a neutral model, a local reduction in levels of variation may be observed due to the large evolutionary variance in the time to the most recent common ancestor. The difficulty lies in determining how large a region and how great of a reduction in levels of variation cannot be explained by a neutral model. Kim and Stephan [31] have developed a maximum likelihood method to test for hitchhiking based on polymorphism sampled along a chromosome. The test is based on both a reduction in levels of variation and a skew in the frequency spectrum.

## Skew in the frequency spectrum

The effect of hitchhiking on the frequency spectrum, depends on the ratio of the recombination rate to the selection coefficient, the initial frequency of the advantageous mutation, and most importantly the time since the start (or end) of the hitchhiking event. During the spread of an advantageous mutation, neutral mutations are swept to either low or high frequency depending on their original linkage relationship with the advantageous mutation. In the absence of recombination, a partial hitchhiking event, or one where the advantageous mutation does not reach fixation, can be detected by a single mutation or haplotype present at a much higher frequency than expected under a neutral model (see below). If there is no recombination and hitchhiking is complete, all variation is removed from a locus.

A skew in the frequency spectrum is also produced as an indirect byprod-

uct of removing all variation from a locus. Subsequent to hitchhiking, new mutations accumulate at low frequency in a population and it is some time before they drift to intermediate or high frequencies. This skew in the frequency spectrum towards low frequency variation can be measured by Tajima's D statistic [49]. Tajima's D is the difference between two estimators of the population parameter  $\theta$  divided by the standard deviation of the difference. Under the Wright-Fisher model the expectation of  $\theta$  is equal to  $4N\mu$ , where  $N$  is effective population size and  $\mu$  is the mutation rate. The two estimators are

$$\hat{\theta}_{\pi} = \sum_{n=1}^{n-1} \frac{2S_i i(n-i)}{n(n-1)} \quad (1)$$

which is based on the average heterozygosity [47] and

$$\hat{\theta}_W = \sum_{n=1}^{n-1} S_i \left( \sum_{n=1}^{n-1} \frac{1}{i} \right)^{-1} \quad (2)$$

which is based on the number of segregating sites divided by a constant, which depends on the sample size  $n$  [53].  $\hat{\theta}_{\pi}$  is most sensitive to intermediate frequency variation, whereas  $\hat{\theta}_W$  is most sensitive to rare (low or high frequency) variation. The reasoning is as follows: a single segregating site at intermediate frequency adds  $10 \times (20-10)/(20 \times 19) = 0.26$  to an estimate of  $\hat{\theta}_{\pi}$  whereas a low frequency variant adds much less  $1 \times (20-1)/380 = 0.05$ . In contrast each segregating site contributes equally to  $\hat{\theta}_W$ . Since most variation in a population is found at low frequencies  $\hat{\theta}_W$  is easily influenced by changes in the number of low frequency variants.

Under neutrality the means of two estimators are expected to be equal to one another. Subsequent to a hitchhiking event that has removed all variation  $\hat{\theta}_W$  is expected to be greater than  $\hat{\theta}_{\pi}$  until new mutations reach intermediate frequency in a population. Simulation studies of hitchhiking events have shown that Tajima's D has quite a bit of power to detect a strong hitchhiking event  $0.2N$  generations subsequent to the fixation of an advantageous mutation [43]. The advantage of this test is that no assumptions are made about how much variation is expected in a population. The disadvantage of this test, as well as all other tests that use polymorphism data, is that while recombination doesn't affect the mean it does affect the variance of the frequency spectrum and test statistics based on the frequency spectrum. Recombination decreases the variance since it enables different mutations

within a sample to have different genealogies. While the rate of recombination can be either measured in the lab or estimated from polymorphism data, these estimates rely on a number of assumptions and often have large confidence intervals [3]. The practical solution that is most often taken is to conservatively assume no recombination in the generation of the cutoff values for a test statistic, or to use a conservative estimate of the recombination rate, typically the lower bound estimate.

A number of other tests, besides Tajima's D, have been developed to detect hitchhiking based on a skew in the frequency spectrum. Fu and Li's  $D_{FL}$  and  $D^*_{FL}$ , test for a difference between  $\hat{\theta}_\pi$  and  $\theta$  estimated from the number of singletons (those mutations found only once in a sample). For  $D^*_{FL}$ , an outgroup is used to distinguish when the derived mutation is found once or  $n - 1$  times in a sample of  $n$ . To provide a general framework in which to compare the frequency spectrum to the neutral expectation Fu derived an estimate of  $\theta$  for every frequency class in a sample;  $\hat{\theta}_i = iS_i$  [19]. Thus, it is now possible to compare any part of the frequency spectrum to the neutral expectation. Comparison of these frequency based tests to one another showed that Tajima's D has the most power to detect a hitchhiking event in the absence of recombination [20].

In the presence of recombination hitchhiking produces a skew in the frequency spectrum quite different from that in the absence of recombination. In the presence of recombination a neutral variant will increase or decrease in frequency depending on whether it is on the same haplotype as the advantageous mutation or not. For a deterministic hitchhiking event, the expected frequency to which it goes depends on the ratio of the rate of recombination to the selection coefficient and the initial frequency of the advantageous mutation [36]. The end result is that subsequent to a strong hitchhiking event neutral variation that has recombined onto the advantageous haplotype is found at either high or low frequencies and forms a bipartite frequency spectrum (Figure 2) [15]. High and low frequency variation refer to the frequency of the derived variant (or new mutation) which is distinguished from the ancestral variant using an outgroup. Subsequent to the hitchhiking event, high frequency variants are lost and new mutations at low frequency accumulate [30] [14] [42].

The bipartite frequency spectrum produced in the presence of recombination can be detected by Tajima's D statistic [15], or any other statistic that measures a differences between rare and common variation. However, low frequency variation is easily influenced by changes in population size

and background selection (see below). On the other hand, an excess of high compared to common frequency variation cannot easily be produced by demographic scenarios (see below).  $\hat{\theta}_H$  is a measure of high frequency variation and is based on the homozygosity of the derived variant.

$$\hat{\theta}_H = \sum_{n=1}^{n-1} \frac{2S_i i^2}{n(n-1)} \quad (3)$$

The H test is the difference between  $\hat{\theta}_\pi$  and  $\hat{\theta}_H$ , and is therefore a test for an excess of high compared to intermediate frequency mutations [15]. Because an outgroup must be used to distinguish high and low frequency mutations, the probability of mis-inference must be incorporated into applications of the H test. The derived state can be mis-inferred if a back-mutation occurs at a site. If all sites have the same mutation rate and thus the same probability of a back-mutation, the probability of mis-inference can be estimated by  $d/3$ , where  $d$  is the rate of divergence corrected for multiple hits and  $1/3$  is the probability a mutation is a back-mutation, A to T, rather than A to G, when A and T are segregating [15]. Differences in the rate of transitions and transversions or other mutational heterogeneities can also be incorporated [15]

Both Tajima's D and the H test have good power to detect hitchhiking in the presence of recombination (Figure 3). In contrast to D the power of H drops rapidly after the hitchhiking event since high frequency variants as measured by  $\hat{\theta}_H$  are readily lost due to drift [30] [14] [42]. Tajima's D retains power for much longer due to the influx of new low frequency variation during the recovery from a hitchhiking event (Figure 3). Because variation is recovered first at low, then intermediate, and then high frequencies, a test for a lack of high frequency variation may retain the most power for the longest period of time subsequent to a hitchhiking event. The difference between  $\hat{\theta}_H$  and  $\hat{\theta}_W$ ,  $H_L$ , is a measure of high compared to low frequency variation and retains power for the longest period of time subsequent to hitchhiking (Figure 4). This can be explained by  $\hat{\theta}_H$  being the last of three estimators of  $\theta$  to reach equilibrium and  $\hat{\theta}_W$  being the first.

Using the expected reduction in heterozygosity in combination with the expected skew in the frequency spectrum in the presence of recombination, Kim and Stephan [31] have implemented a maximum likelihood approach to simultaneously test for hitchhiking and then estimate both the location of the advantageous substitution and the strength of selection, given the recombination rate. Although this test appears more powerful than those tests

based on different estimators of  $\theta$ , the test requires precise knowledge of the recombination rate and may be more sensitive to non-equilibrium conditions, since the null and alternative hypothesis are more precisely specified. Yet, it should be noted that the robustness of all tests to violations of the assumptions of the Wright-Fisher model has not been well characterized (see below). In one of the first attempts to explicitly test selective versus demographic explanations, Galtier *et al.* [21] have used a maximum likelihood approach to distinguish selection from a population bottleneck using data from *Drosophila* for which multiple loci have been surveyed for polymorphism. The logic behind the test is that a population bottleneck is expected to reduce levels of variation and skew the frequency spectrum across all loci whereas a hitchhiking event is expected to be specific to a fraction of loci.

## Linkage disequilibrium

Hitchhiking is expected to produce linkage disequilibrium both in the presence and absence of recombination [50]. During the spread of an advantageous mutation through a population, a haplotype of very tightly linked neutral variants will increase in frequency until fixation. In some instances a second haplotype may remain segregating at appreciable frequencies ( $> 1\%$ ) by recombining onto the advantageous chromosome during the hitchhiking event. Farther away from the site under selection recombination events allow one or more different haplotypes to recombine onto the advantageous chromosome and escape loss. As the distance to the site under selection increases so do the number of alleles that escape complete hitchhiking (Figure 3 of [15]). If the rate of recombination is low enough so that there is no recombination within the sequence surveyed but enough recombination so that variation remains segregating subsequent to hitchhiking, a strong haplotype pattern may form where all variation is divided among only a few haplotypes. In the extreme case where only two haplotypes remain segregating, all variation may be in complete linkage disequilibrium. A neutral model may not be able to explain the presence of a single haplotype at intermediate or high frequency [25] [33]. In addition to hitchhiking with recombination, a single haplotype could reach high frequency due to balancing selection, the loss of positive selection during a hitchhiking event, or interference with advantageous or deleterious mutations in the population [25] [33]. The degree to which hitchhiking produces linkage disequilibrium between two alleles can

be measured by  $r$ , their correlation coefficient and  $D'$  the difference between the observed and expected (assuming independence) biallelic frequencies in a sample [35].

$$r = \frac{f_{AB} - f_A f_B}{\sqrt{f_A f_B (1 - f_A)(1 - f_B)}} \quad (4)$$

$$D' = \frac{f_{AB} - f_A f_B}{\min[f_A f_B, (1 - f_A)(1 - f_B)]} \text{ for } D' > 0 \quad (5)$$

$$D' = \frac{f_{AB} - f_A f_B}{\min[f_A(1 - f_B), (1 - f_A)f_B]} \text{ for } D' < 0 \quad (6)$$

where  $f_A$  is the frequency of the major allele at the first locus,  $f_B$  is the frequency of the major allele at the second locus and  $f_{AB}$  is the frequency of the AB haplotype. Strong hitchhiking produces more linkage disequilibrium than expected in the absence of recombination, when measured by  $r$  and  $D'$  [14] [31] [42]. This is true even when some recombination is allowed between the two neutral markers during hitchhiking (Figure 5). However, previous work has shown that linkage disequilibrium decays rapidly subsequent to hitchhiking [42]. More work is necessary to distinguish linkage disequilibrium created by demographic effects or selection.

A number of haplotype tests have been developed to detect a high frequency haplotype or a lack of haplotype diversity that may occur during or subsequent to a hitchhiking event. Hudson *et al.* [25] developed a test to determine the probability of observing a given number of segregating sites or fewer in a subset of sequences from a sample, and applied this to the *Sod* locus. The  $F_s$  test [20], is equal to  $\ln(S/(1 - S))$ , where  $S$  is the probability of having no fewer than  $k$  alleles in a sample given  $\hat{\theta}_\pi$  [12]. Depaulis and Veuille [12] have proposed two tests for an excess of linkage disequilibrium.  $H_{DV}$  is the observed haplotype diversity,  $K$  is the number of haplotypes, and both are conditioned on the number of segregating sites in a sample.  $K$  and  $F_s$  are only different in that they are conditioned on different estimators of  $\theta$ .

## Population subdivision and changes in population size

The effect of hitchhiking on linked neutral variation in a structured population or one that has recently changed in size is not easily understood.

However, in most cases the qualitative dynamics of hitchhiking are expected to be the same; variation is removed from a population producing a skew in the frequency spectrum and linkage disequilibrium. Hitchhiking in a structured population is particularly difficult to describe since it depends on the number of subpopulations, the migration rates between these populations, and the effective size of these subpopulations. When the number of emigrants is less than one per generation it has been shown that hitchhiking produces population differentiation as a function of the strength of selection [44]. The effect of hitchhiking in a two dimensional model of isolation by distance has also been studied [6].

More important than understanding how hitchhiking is affected by population structure or changes in population size, is how the assumption of a constant panmictic population affects current methods of detecting hitchhiking. If demographic forces produce patterns that resemble hitchhiking, then the rate of erroneously detecting a hitchhiking event when none has occurred (rate of false positives) may be high. If demographic forces produce a pattern opposite to that of hitchhiking, then the power of detecting hitchhiking (rate of true positives) may be low. For all of the above mentioned tests, the rate of true and false positives is affected by both population subdivision and changes in population size. This results both from the effect of demography on the expectation of statistics such as Tajima's  $D$ , but also from the effect of demography on the variance in  $D$ . Selective forces are often distinguished from demographic forces since the former is expected to be locus specific while the latter is expected to affect the entire genome. However, if demography has a slight effect on the mean value of a test statistic or only affects the variance of a test statistic it is likely to go unnoticed if only a few loci across the genome are examined. Thus, it is important to know how changes in population size and population subdivision affect various tests used to detect hitchhiking.

A change in population size affects levels of variation, the frequency spectrum and linkage disequilibrium. An increase in population size causes an increase in levels of low frequency variation and a negative Tajima's  $D$  value whereas a decrease in population size causes a decrease in levels of low and high frequency variation and a positive Tajima's  $D$  value [48]. The variance in Tajima's  $D$  has been shown to decrease in an expanding population [40] and is likely increased in a shrinking population. An increase in population size also causes a decrease in linkage disequilibrium as measured by  $r$  [41].

Population structure affects patterns of variation in a much more compli-

cated way. Tajima [48] studied a simple model of two demes with balanced migration. In the case where samples are drawn from both subpopulations, the heterozygosity increases faster than the number of segregating sites with decreasing rates of migration, producing positive Tajima's D values. If samples are drawn from just one of the subpopulations, heterozygosity remains unchanged while the number of segregating sites decreases slightly with intermediate rates of migration  $4Nm \approx 1$ , producing slightly negative Tajima's D values. In contrast, when the rate of migration is 19 times greater from one population to the other and samples are drawn from both subpopulations, the number of segregating sites increases faster than heterozygosity with decreasing rates of migration. Wakeley [51] found the variance in heterozygosity both within and between populations increases with migration rate for a two subpopulation model with balanced migration. Population subdivision is also known to increase levels of linkage disequilibrium [52].

Although few statistics have been tested for sensitivity to different population histories, there are obvious cases in which a population's history can appear similar to hitchhiking. For Tajima's D and Fu and Li's  $D_{FL}$  this is a recent increase in population size, for the H test this is the presence of a rare migrant from a distantly related population or species, for the haplotype based tests this is population subdivision or recent admixture. One case has been studied for Tajima's D and the H test. For a two subpopulation model with balanced migration where 50 alleles are sampled from a single subpopulation, Tajima's D is significant in 6% and 9% of cases for  $4Nm = 1$  and  $4Nm = 0.5$ , respectively, whereas the H test is significant in 14% and 19% of cases for  $4Nm = 1$  and 0.5, respectively [42]. However, under most circumstances the D and H tests would be applied to a genetically diverse sample. Because subdivision tends to produce an excess of intermediate compared to low frequency variation when sample are drawn from a mixture of subpopulations, the D and H statistics are likely conservative.

The simplest way of distinguishing demographic from selective explanations is by surveying other unlinked loci in the genome. Any demographic explanation is expected to affect all loci whereas selection is expected to be specific to only a few loci. Subtle demographic effects, such as an increase in the variance of a statistic, are the most worrisome since they may go unnoticed in a survey of a small number of genes but may affect the rate of false positives of a test. Multiple independent lines of evidence, such as a regional reduction in levels of variation in combination with a skew in the frequency spectrum should be used to rule out a demographic explanation.

## Distinguishing background selection and hitchhiking in regions of low recombination

One of the few genome wide patterns in polymorphism data that cannot be attributed to mutation and drift is the correlation between levels of variation and rates of recombination. This observation has now been made in numerous species, but it is still debated as to its cause [2]. The observation cannot be explained by different mutation rates since rates of recombination are not correlated with divergence between species. However, there are now many examples of heterogeneity in levels of divergence between two species suggesting mutation rates may vary across the genome [56]. A question that has not been answered is the extent to which heterogeneity in levels of variation across the genome can be explained by mutational heterogeneities alone. The effect of regional differences in mutation rates across the genome must be accounted for in explaining low levels of variation in regions of low recombination.

Both background selection and recurrent hitchhiking can produce reduced levels of variation in regions of low recombination. With a sufficiently high rate of deleterious mutations per cM, background or purifying selection against deleterious mutations removes linked neutral variation, essentially reducing a population's effective size [10]. With a sufficiently high rate of adaptive substitutions driven by sufficiently strong selection, recurrent hitchhiking events may also maintain low levels of variation across an entire region of low recombination [9]. Tajima's D statistic is often used to distinguish between background selection and hitchhiking [4]. Simulation studies have shown that recurrent hitchhiking events in the presence of recombination produce an excess of low frequency variants and significantly negative D values [9]. In contrast, simulation studies have shown that background selection produces little or no skew in the frequency spectrum if  $Ns$  is sufficiently large, where  $s$  is the strength of selection against deleterious mutations [10] [11] [20]. When background selection affects the frequency spectrum, Fu and Li's D has the most power to detect it [20]. Numerous polymorphism surveys were conducted in regions of low recombination with the aim of distinguishing background selection from hitchhiking by means of a skew in the frequency spectrum as measured by Tajima's D [8] [7] [24] [54] [34] [4] [28]. However, in many of these cases there was so little variation found that there was no power to detect a significant skew in the frequency

spectrum.

If selection is so weak that deleterious mutations reach detectable frequencies ( $> 1\%$ ) in a population, these mutations and neutral mutations linked to them are expected to produce an excess of low compared to common frequency variation. Studies of allozyme variation in humans and fruit flies indicate that a large proportion of low frequency amino acid variants are slightly deleterious and reach detectable frequencies in a population [39]. By comparing the distribution of amino acid to synonymous variation demographic explanations were ruled out and many of these deleterious mutations were shown to reach frequencies of 1-10% for both humans [16] and *D. melanogaster* [17]. Forward simulations of purifying selection show that mutations with  $2Ns$  values as small as 6 can reduce levels of variation and produce negative  $D$  values in the absence of recombination [23]. The same effect is found when deleterious mutations are gamma distributed and there is no recombination [57]. Thus, at least in the absence of recombination, background selection as well as hitchhiking may produce negative  $D$  values if a sufficient number of deleterious mutations are slight in their effects.

The  $H$  test can be used to distinguish hitchhiking and background selection in regions of low recombination. The  $H$  statistic should not be affected by background selection, which only skews the frequency spectrum at low frequencies. In fact, in the presence of background selection hitchhiking may produce more high compared to intermediate frequency variants than in the absence of background selection. The greater number of high frequency variants is the result of the excess of low frequency variants present prior to hitchhiking. It is these low frequency variants that are swept to high frequencies during hitchhiking. Thus, under the extreme example where only low frequency variants are present in a population, hitchhiking may produce only high frequency variants since all low frequency variants are either swept to high frequency or to frequencies too low to be detected. There are a number of regions where this has been observed. The y-ac region is on the tip of the X chromosome of *D. melanogaster* and shows three high frequency restriction sites [15]. Five olfactory receptor pseudogenes in a 450kb region in humans contain predominantly high frequency variants [22].

To distinguish background selection from hitchhiking the  $H$  test must have reasonable power to detect recurrent hitchhiking events. Recurrent hitchhiking is different from a single hitchhiking event since at the start of each hitchhiking event the population is not at equilibrium. In most instances the population is likely recovering from the last hitchhiking event and so

should have an excess of low frequency variants. The next hitchhiking event is expected to sweep low frequency variation to high or lower frequencies. Although coalescence simulations of recurrent hitchhiking events show the H test has little power to detect recurrent hitchhiking events, this has been shown only for very strong selection and infrequent hitchhiking events, a limitation of the approach [42]. Under these conditions the power of detecting hitchhiking using the H test drops quickly subsequent to the fixation of the advantageous mutation. However, as the frequency of hitchhiking events increases the neutral frequency spectrum may approach a U shaped distribution, which is the expected frequency distribution for mutations under positive selection [13].

Finally, background selection and hitchhiking may be distinguished in a subdivided population if hitchhiking reduces variation in only one of the subpopulation or much more in one of the subpopulations [46]. Using a reference locus as a control for the expected reduction in levels of variation due to background selection, the *vermillion* locus was shown to have significantly reduced levels of variation in two of four subpopulations of *D. ananassae* [46].

## Conclusions and future directions

Significant advances have been made in how positive selection is detected using DNA polymorphism data. While a slew of new test statistics have been developed and shown to have power to detect hitchhiking, it is standard practice to assume no recombination and a randomly mating Wright-Fisher population in determining the cutoff values for these tests. As genomic surveys of polymorphism become available, reliable estimates of the recombination rate and populations' demographic history can be made [18] [40]. In the meantime convincing evidence for hitchhiking must include multiple lines of evidence such as a local reduction in levels of variation and a local skew in the frequency spectrum.

Genomic surveys of polymorphism will provide some indication of the number and location of loci in the genome that have recently experienced a hitchhiking event and the relative contributions of background selection and hitchhiking to the reduction in levels of variation in regions of low recombination. This can only be done by examination of high frequency variation since low frequency variation is similarly influenced by both background selection and hitchhiking.

## References

- [1] M Aguade, N Miyashita, and CH Langley. Polymorphism and divergence in the Mst26A male accessory gland gene region in *Drosophila*. *Genetics*, 132(3):755–770, 1992.
- [2] P Andolfatto. Adaptive hitchhiking effects on genome variability. *Curr Opin Genet Dev*, 11(6):635–641, 2001.
- [3] P Andolfatto and M Przeworski. A genome-wide departure from the standard neutral model in natural populations of *Drosophila*. *Genetics*, 156(1):257–268, 2000.
- [4] P Andolfatto and M Przeworski. Regions of lower crossing over harbor more rare variants in African populations of *Drosophila melanogaster*. *Genetics*, 158(2):657–665, 2001.
- [5] NH Barton. The effect of hitch-hiking on neutral genealogies. *Genet Res*, 72:123–133, 1998.
- [6] NH Barton. Genetic hitchhiking. *Philos Trans R Soc Lond B Biol Sci*, 355(1403):1553–1562, 2000.
- [7] DJ Begun and CF Aquadro. Evolution at the tip and base of the X chromosome in an African population of *Drosophila melanogaster*. *Mol Biol Evol*, 12(3):382–390, 1995.
- [8] AJ Berry, JW Ajioka, and M Kreitman. Lack of polymorphism on the *Drosophila* fourth chromosome resulting from selection. *Genetics*, 129(4):1111–1117, 1991.
- [9] JM Braverman, RR Hudson, NL Kaplan, CH Langley, and W Stephan. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics*, 140(2):783–796, 1995.
- [10] B Charlesworth, MT Morgan, and D Charlesworth. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4):1289–1303, 1993.
- [11] D Charlesworth, B Charlesworth, and MT Morgan. The pattern of neutral molecular variation under the background selection model. *Genetics*, 141(4):1619–1632, 1995.

- [12] F Depaulis and M Veuille. Neutrality tests based on the distribution of haplotypes under an infinite-site model. *Mol Biol Evol*, 15(12):1788–1790, 1998.
- [13] WJ Ewens. *Mathematical population genetics*. Springer-Verlag, 1979.
- [14] JC Fay. *Detecting natural selection from patterns of DNA polymorphism and divergence*. PhD thesis, University of Chicago, 2001.
- [15] JC Fay and CI Wu. Hitchhiking under positive Darwinian selection. *Genetics*, 155(3):1405–1413, 2000.
- [16] JC Fay, GJ Wyckoff, and CI Wu. Positive and negative selection on the human genome. *Genetics*, 158(3):1227–1234, 2001.
- [17] JC Fay, GJ Wyckoff, and CI Wu. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature*, 415(6875):1024–1026, 2002.
- [18] L Frisse, RR Hudson, A Bartoszewicz, JD Wall, J Donfack, and Rienzo A Di. Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *Am J Hum Genet*, 69(4):831–843, 2001.
- [19] YX Fu. Statistical properties of segregating sites. *Theor Popul Biol*, 48(2):172–197, 1995.
- [20] YX Fu. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics*, 147(2):915–925, 1997.
- [21] N Galtier, F Depaulis, and NH Barton. Detecting bottlenecks and selective sweeps from DNA sequence polymorphism. *Genetics*, 155(2):981–987, 2000.
- [22] Y Gilad, D Segre, K Skorecki, MW Nachman, D Lancet, and D Sharon. Dichotomy of single-nucleotide polymorphism haplotypes in olfactory receptor genes and pseudogenes. *Nat Genet*, 26(2):221–224, 2000.
- [23] I Gordo, A Navarro, and B Charlesworth. Muller’s Ratchet and the Pattern of Variation at a Neutral Locus. *Genetics*, 161(2):835–848, 2002.

- [24] MT Hamblin and CF Aquadro. High nucleotide sequence variation in a region of low recombination in *Drosophila simulans* is consistent with the background selection model. *Mol Biol Evol*, 13(8):1133–1140, 1996.
- [25] RR Hudson, K Bailey, D Skarecky, J Kwiatowski, and FJ Ayala. Evidence for positive selection in the superoxide dismutase (Sod) region of *Drosophila melanogaster*. *Genetics*, 136(4):1329–1340, 1994.
- [26] RR Hudson, M Kreitman, and M Aguade. A test of neutral molecular evolution based on nucleotide data. *Genetics*, 116(1):153–159, 1987.
- [27] RR Hudson, AG Saez, and FJ Ayala. DNA variation at the Sod locus of *Drosophila melanogaster*: an unfolding story of natural selection. *Proc Natl Acad Sci U S A*, 94(15):7725–7729, 1997.
- [28] MA Jensen, B Charlesworth, and M Kreitman. Patterns of genetic variation at a chromosome 4 locus of *Drosophila melanogaster* and *D. simulans*. *Genetics*, 160(2):493–507, 2002.
- [29] NL Kaplan, RR Hudson, and CH Langley. The "hitchhiking effect" revisited. *Genetics*, 123(4):887–899, 1989.
- [30] Y Kim and W Stephan. Joint effects of genetic hitchhiking and background selection on neutral variation. *Genetics*, 155(3):1415–1427, 2000.
- [31] Y Kim and W Stephan. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*, 160(2):765–777, 2002.
- [32] M Kimura and T Ota. The average number of generations until extinction of an individual mutant gene in a finite population. *Genetics*, 63(3):701–709, 1969.
- [33] DA Kirby and W Stephan. Haplotype test reveals departure from neutrality in a segment of the white gene of *Drosophila melanogaster*. *Genetics*, 141(4):1483–1490, 1995.
- [34] CH Langley, BP Lazzaro, W Phillips, E Heikkinen, and JM Braverman. Linkage disequilibria and the site frequency spectra in the su(s) and su(w(a)) regions of the *Drosophila melanogaster* X chromosome. *Genetics*, 156(4):1837–1852, 2000.

- [35] RC Lewontin. The interaction of selection and linkage. I. General considerations heterotic models. *Genetics*, 49:49–67, 1964.
- [36] J Maynard-Smith and J Haigh. The hitch-hiking effect of a favourable gene. *Genet Res*, 23(1):23–35, 1974.
- [37] EN Moriyama and JR Powell. Intraspecific nuclear DNA variation in *Drosophila*. *Mol Biol Evol*, 13(1):261–277, 1996.
- [38] D Nurminsky, DD Aguiar, CD Bustamante, and DL Hartl. Chromosomal effects of rapid gene evolution in *Drosophila melanogaster*. *Science*, 291(5501):128–130, 2001.
- [39] T Ohta. Statistical analyses of *Drosophila* and human protein polymorphism. *Proc Natl Acad Sci U S A*, 72:3194–3196, 1975.
- [40] A Pluzhnikov, A Di Rienzo, and RR Hudson. Inferences about human demography based on multilocus analyses of noncoding sequences. *Genetics*, 161(3):1209–1218, 2002.
- [41] JK Pritchard and M Przeworski. Linkage disequilibrium in humans: models and data. *Am J Hum Genet*, 69(1):1–14, 2001.
- [42] M Przeworski. The signature of positive selection at randomly chosen Loci. *Genetics*, 160(3):1179–1189, 2002.
- [43] KL Simonsen, GA Churchill, and CF Aquadro. Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics*, 141(1):413–429, 1995.
- [44] M Slatkin and T Wiehe. Genetic hitch-hiking in a subdivided population. *Genet Res*, 71(2):155–160, 1998.
- [45] W Stephan, THE Wiehe, and MW Lenz. The effect of strongly selected substitutions on neutral polymorphism: Analytical results based on diffusion theory. *Theor Popul Biol*, 41:237–254, 1992.
- [46] W Stephan, L Xing, DA Kirby, and JM Braverman. A test of the background selection hypothesis based on nucleotide data from *Drosophila ananassae*. *Proc Natl Acad Sci U S A*, 95(10):5649–5654, 1998.

- [47] F Tajima. Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105(2):437–460, 1983.
- [48] F Tajima. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3):585–595, 1989.
- [49] F Tajima. The effect of change in population size on DNA polymorphism. *Genetics*, 123(3):597–601, 1989.
- [50] G Thomson. The effect of a selected locus on linked neutral loci. *Genetics*, 85(4):753–788, 1977.
- [51] J Wakeley. The variance of pairwise nucleotide differences in two populations with migration. *Theor Popul Biol*, 49(1):39–57, 1996.
- [52] JD Wall. Detecting ancient admixture in humans using sequence polymorphism data. *Genetics*, 154(3):1271–1279, 2000.
- [53] GA Watterson. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol*, 7(2):256–276, 1975.
- [54] ML Wayne and M Kreitman. Reduced variation at concertina, a heterochromatic locus in *Drosophila*. *Genet Res*, 68(2):101–108, 1996.
- [55] TH Wiehe and W Stephan. Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. *Mol Biol Evol*, 10(4):842–854, 1993.
- [56] EJ Williams and LD Hurst. Is the synonymous substitution rate in mammals gene-specific? *Mol Biol Evol*, 19(8):1395–1398, 2002.
- [57] S Williamson and ME Orive. The genealogy of a sequence subject to purifying selection at multiple sites. *Mol Biol Evol*, 19(8):1376–1384, 2002.

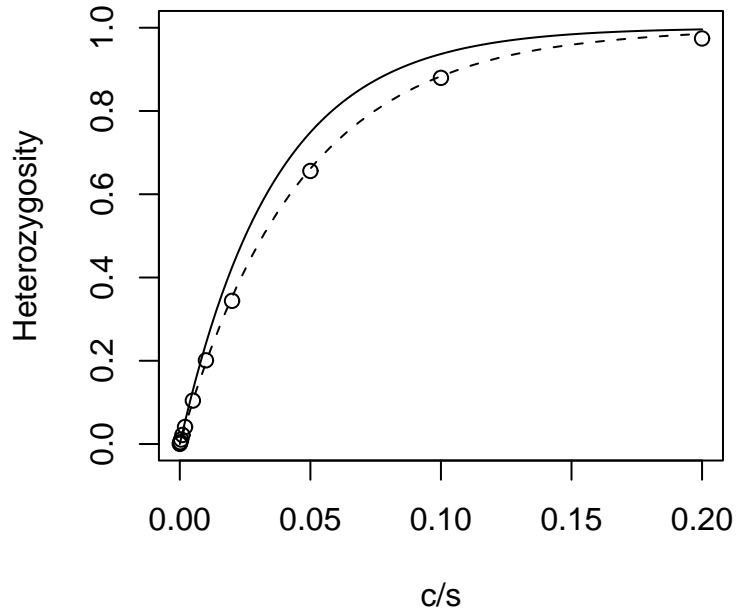


Figure 1: Heterozygosity as a function of  $c/s$  for the deterministic approximation of Maynard Smith and Haigh [36], *eq. 8*  $\approx 1 - e^{-2c/s}$  (solid line), the deterministic approximation of Stephan *et al.* [45], *eq. 17* (dashed line), and for  $10^4$  coalescence simulations (circles). Simulation parameters are  $2N = 10^8$ ,  $s = 10^{-3}$ ,  $\varepsilon = 10^{-6}$  and is the initial frequency of the advantageous mutation.

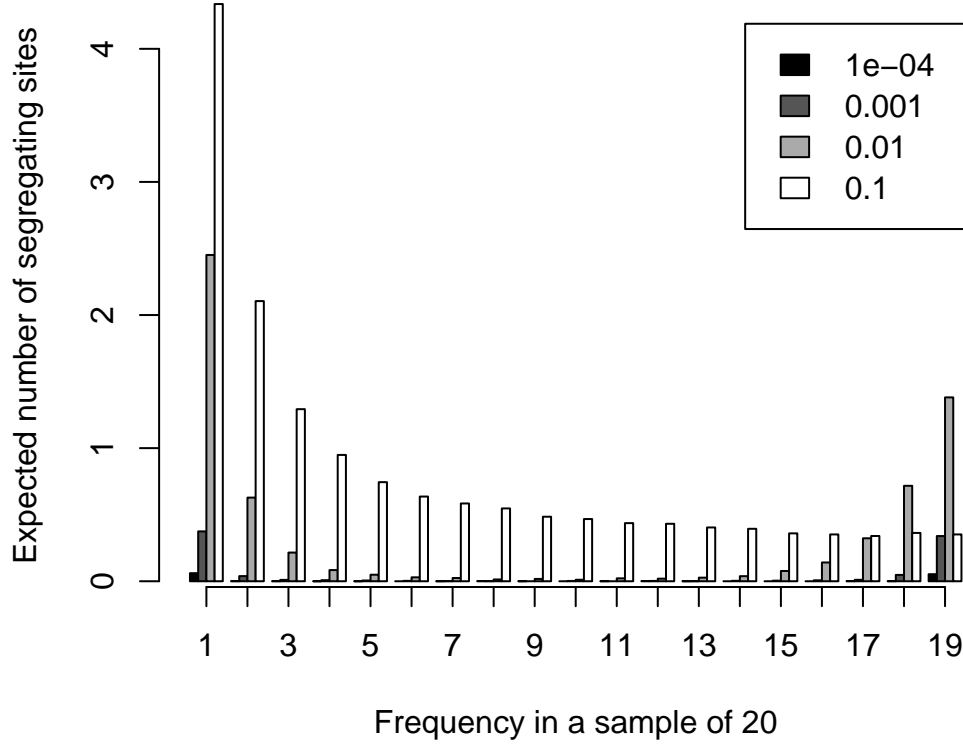


Figure 2: Expected frequency spectrum of sites in a sample of 20 subsequent to a hitchhiking event for different  $c/s$  values. Parameters are  $10^4$  coalescence simulations,  $2N = 10^8$ ,  $s = 10^{-3}$ ,  $\theta = 5$ , sample size is 20.

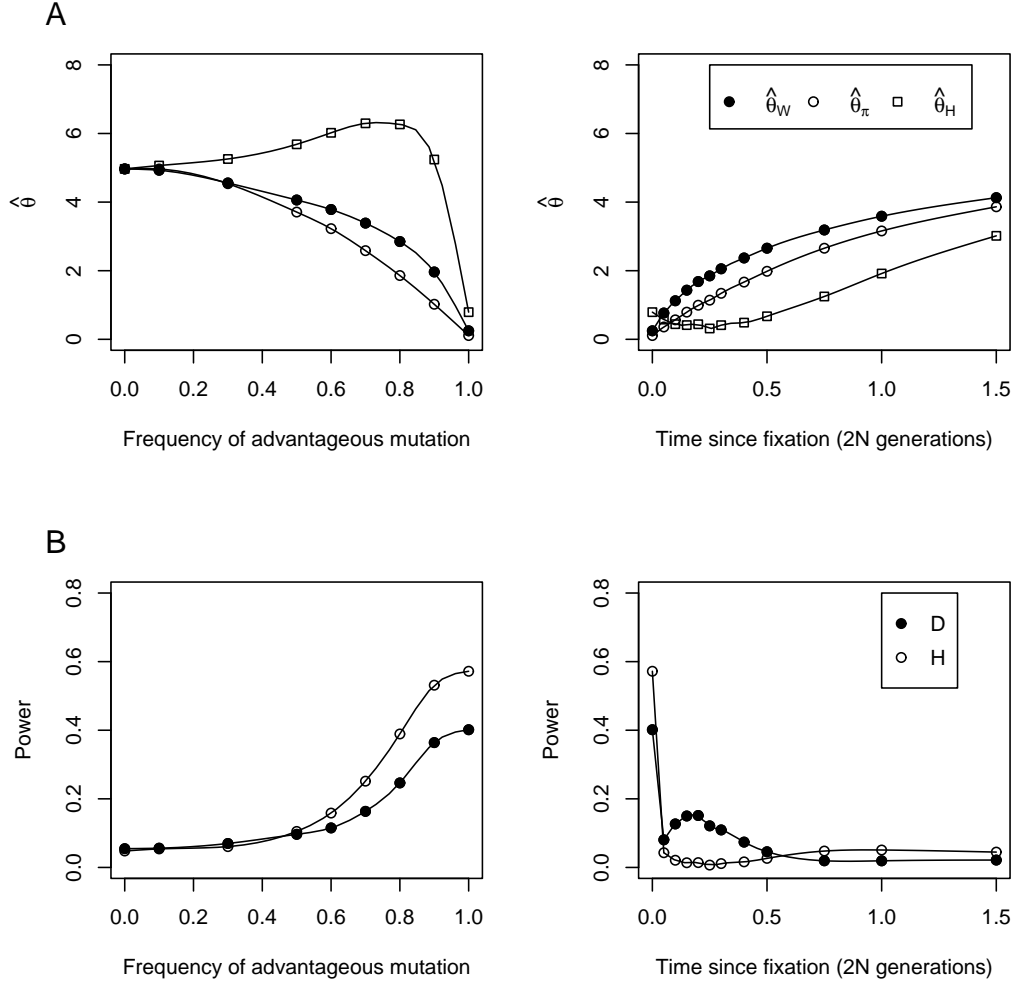


Figure 3: A - The expectation of different estimators of  $\theta$  during and subsequent to hitchhiking. B - The power of the D and H statistics during and subsequent to hitchhiking. The simulation parameters are the same as in Figure 2 except  $c/s$  is fixed at  $10^{-3}$ . For each simulated hitchhiking event with at least one segregating site D and H were compared to critical values generated from  $10^4$  neutral coalescence simulations with a fixed number of segregating sites equal to that observed in the hitchhiking simulation.

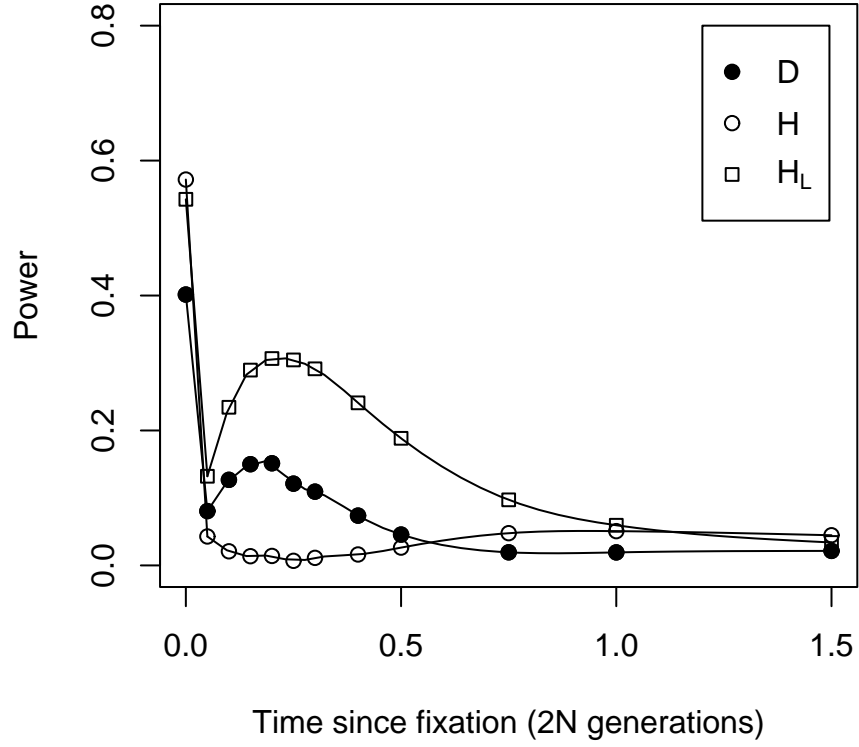


Figure 4: The power of D, H and  $H_L$  as a function of time since hitchhiking.  $H_L$  is the difference between  $\hat{\theta}_W$  and  $\hat{\theta}_H$ . The simulation parameters are the same as those in Figure 3.

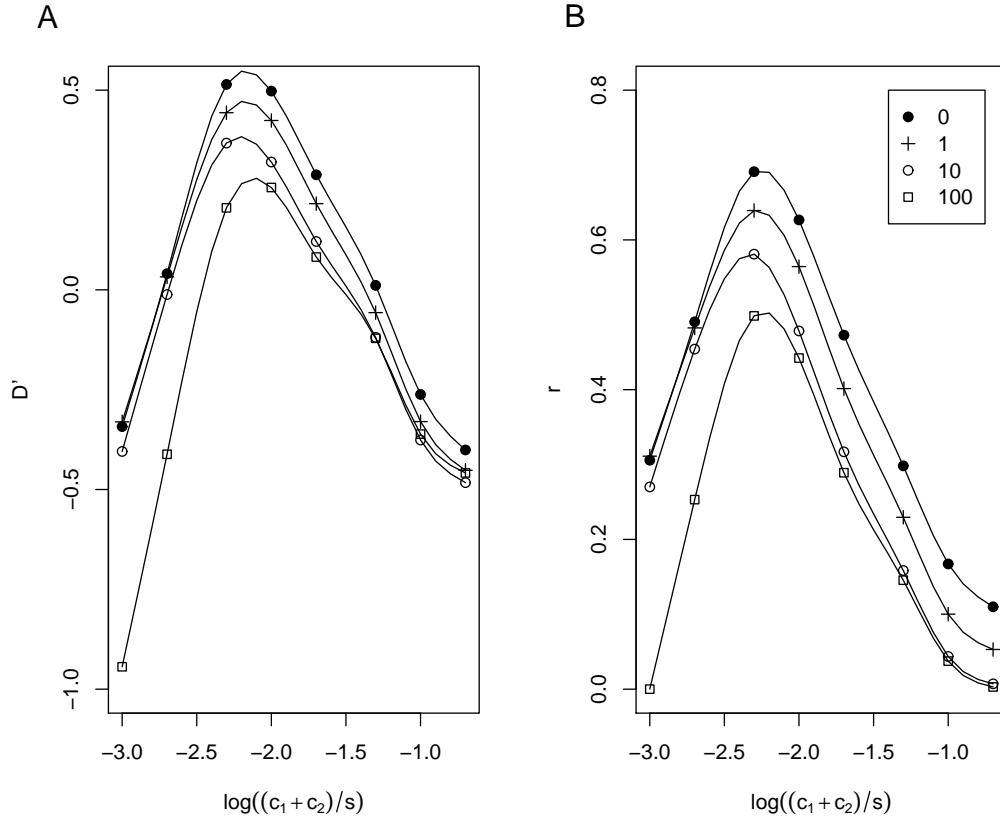


Figure 5: The average of  $r$  (A) and  $D'$  (B) as a function  $(c_1+c_2)/s$ , where  $c_1$  is the rate of recombination between the selected locus and adjacent neutral locus and  $c_2$  is the rate of recombination between the two neutral loci.  $4Nc_2 = 0$  (solid circles),  $4Nc_2 = 1$  (cross),  $4Nc_2 = 10$  (open circles),  $4Nc_2 = 100$  (squares), sample size is 50,  $2N = 10^8$ .